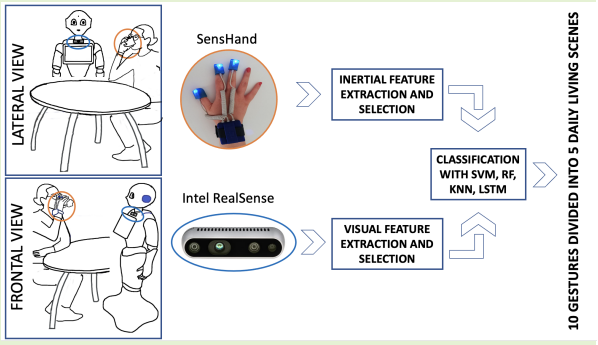
Daily gesture recognition during human-robot interaction combining vision and wearable systems

FIORINI, Laura, LOIZZO, Federica G Cornacchia, SORRENTINO, Alessandra, KIM, Jaeseok, ROVINI, Erika, DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650> and CAVALLO, Filippo

Daily gesture recognition during human-robot interaction combining vision and wearable systems

Laura Fiorini, *Member, IEEE,* Federica G.Cornacchia Loizzo, Alessandra Sorrentino, Jaeseok Kim, Erika Rovini, *Member, IEEE,* Alessandro Di Nuovo, *Senior Member, IEEE,* Filippo Cavallo, *Member, IEEE*

***Abstract*— The recognition of human gestures is crucial for improving the quality of human-robot cooperation. This article presents a system composed of a Pepper robot that mounts an RGB-D camera and an inertial device called SensHand. The system acquired data from twenty people who performed five daily living activities (i.e. Having Lunch, Personal Hygiene, Working, House Cleaning, Relax). The activities were composed of at least two "basic" gestures for a total of 10 gestures. The data acquisition was performed by two cameras positioned laterally and frontally to mimic the real conditions. The acquired data were off-line classified considering different combinations of sensors to evaluate how the sensor fusion approach improves the recognition abilities. Specifically, the article presents an experimental study that evaluated four algorithms often used in computer vision, i.e. three classical machine learning and one belonging to the field of deep learning, namely Support Vector Machine, Random Forest, K-Nearest Neighbor and Long Short-Term Memory Recurrent Neural Network. The comparative analysis of the results shows a significant improvement of the accuracy when fusing camera and sensors data, i.e. 0.81 for the whole system configuration when the robot is in a frontal position with respect to the user (0.79 if we consider only the index finger sensors) and equal to 0.75 when the robot is in a lateral position. Interestingly, the system performs well in recognising the transitions between gestures when these are presented one after the other, a common event in the real-life that was often neglected in the previous studies.**



***Index Terms*— Gesture recognition, Human-robot interaction, Inertial sensor, Social robot**

I. Introduction

Nowadays social robots permeate our daily life such as workplaces and hospitals. Indeed, they are required to naturally interact and cooperate with human-beings to recognize what a person is doing in a particular moment of the day [1]. In this case, they should be able to sense and perceive what it is happening around them and proactively operate [2].

Human beings' communication is made of verbal and non verbal cues. Among the non verbal ones, we can find emotions, gestures, body postures and gazes. Gestures are even more important in noisy environments, at a distance, and for people with hearing impairments. According to [3], we can distinguish three types of gestures: i) *Body gestures* that refer to the full-body actions or motions; ii) *Hand and arm gestures* that include arm poses and hand gestures; iii) *Head and facial gestures* that refer to nodding or shaking head such as to winking lips.

Liu et al. [4] proposed a framework for the recognition of human gestures for robotic collaboration which is composed of five essential parts: sensor data collection, gesture identification, gesture tracking, gesture classification and gesture mapping, which is the translation of the recognized gestures into an action command for the robot. For what concerns the gesture identification and tracking, it is important to find the right combination of sensors that could track the gesture in case of occlusion, reduced light conditions and comfort to use the sensors (e.g. privacy, wearability) above all the issues [5].

As for the gesture classification, in a real case scenario, simple human hand gestures are usually part of complex activities, more difficult to recognize, that could also include the

Research was pursued with the co-founding of European Union - FESR o FSE, PON Research and Innovation 2014-2020 Project ARS01_01120, SI-ROBOTICS – "Healthy and active ageing through Social ROBOTICS"

L. Fiorini and F. Cavallo are with the Department of Industrial Engineering, University of Florence, Florence, Italy (Corresponding Author: Laura Fiorini, laura.fiorini@unifi.it)

F.G. Cornacchia Loizzo, A. Sorrentino, J. Kim, E. Rovini and F.Cavallo are with the BioRobotics Institute, Scuola Superiore Sant'Anna, 56025 Pontedera, Pisa, Italy and with the Department of Excellence in Robotics & AI, 56127 Pisa, Italy (e-mail: name.surname@santannapisa.it).

A. Di Nuovo is with Sheffield Robotics and the Department of Computing, Sheffield Hallam University, Sheffield, United Kingdom (e-mail: a.dinuovo@shu.ac.uk).

movement of the body such as different poses. For instance, the reader can consider the list of "basic gestures" that you are performing when you are pealing an apple or working at your desk. For this reason, it is very important that social robots gain the ability to distinguish these simple gestures even when they are part of daily living “scenes”.

Therefore, our work aims at developing a multi-modal system in which inertial and visual data are combined together to offer robust human gesture recognition during a daily living scene. Skeleton data were obtained from an RGB-D camera mounted over a social humanoid robot, Pepper, and they were combined with the inertial data acquired by a wearable device, SensHand, that is able to acquire inertial data from the index finger and the wrist. In this approach, the information about the fine movements was collected with the wearable device even when the person was not in the optimal field of view of the robot (i.e. lateral position). On the other hand, the use of the depth camera mounted on the robot gives information about the whole body posture, increasing thus the ability to distinguish among different activities that could have similar hand gestures.

In this paper, we evaluate the performances of the system in realistic cases, specifically investigating if it can reliably perform the gesture recognition when the subjects perform some daily living scenes, arbitrarily alternating gestures and body postures. Since one of the main problems of vision-based sensors is related to the camera occlusion, in this work we evaluate the performances of the system in two real-case scenarios, i.e. when the robot is in front of the person and when it is on the side.

II. Related Works

According to the recent review papers [4][17][18], the most commonly used sensors in the gesture recognition applications are the RGB-D video cameras and the inertial wearable sensors. The former are widely available and cost effective. They provide a rich texture information of the scene and they are easy to operate. However, they have some limitations related to background clutter, occlusion, camera position, subject variations in performing the actions and they are limited to a constrained space defined by the camera position and settings [5]. On the contrary, inertial sensors enable coping with a much wider field of view as well as changing lighting conditions. Thanks to the decrease in the energy consumption and the increase in the computational power of the inertial sensors, long-term recordings have been enabled over the last years. Indeed, many authors focused on the use of this kind of sensors to perform human activity recognition. However, inertial sensors have limitations as well. One of the main restrictions is the sensor drift that may occur during long operational times; moreover, measurements are sensitive to sensor location on the body. In addition, wearable sensors for human action recognition require to be worn by subjects performing the actions, which create the disadvantage of intrusiveness or inconvenience for the subjects [19]. It is also evident that no single sensor modality can cope with various situations that may occur in real scenarios. For example, think about all the actions you are performing while you are cooking in your kitchen or you are brushing your teeth. Some of these actions may be very similar from body/arm movements point of view, but they could differ from hand fine movements point of view (i.e fingers movement). Drink from a glass or talking to the phone have the same arm movement (i.e. bring the arm towards the face) but a different movement and positions of wrist and fingers. In this sense, one way to improve the performance of the human action recognition systems is to combine data from these two different modality sensors considering that they provide complementary information [5] [4].

In this context, several works focused on the use of multimodal sensors to perform activity recognition. Particularly, some of them focused on the recognition of basic body positions and movements (i.e. walking, running, lying) [15], [7], [8], [10], [16], rather than more complex activities of daily living. Other gesture recognition papers, such as public datasets, selected activities of daily living in addition to the basic actions [13], [6], [9], [11]. However, they use data acquired by cameras and/or inertial units placed on the wrist or chest, focusing more on the recognition of the full body activities rather than the recognition of fine finger gestures, even if hand gestures play a pivotal role within actions of daily living, as showed in [14]. Another limitation of the literature works relies on the "scene" composition. Indeed, most of the literature works and public data acquire data from controlled settings [13] [16], where participants just performed one "basic" activity at a time in front of the camera. Nevertheless, this is not a realistic situation, since most of the "scene" we’re daily performing are composed by basic gestures. Moreover, the subject may not be positioned in a frontal position with respect to the camera, but in a lateral one, causing the visual occlusion of the body joints and affecting the accuracy of the system. Table I summarizes and compares the related works in terms of type of tracked movements (i.e. wide movements that involve the whole body or fine movements of the hand), the activities mapping (i.e. basic, activities of daily living, scene) and the type of classifiers (i.e. machine learning or deep learning). Therefore, in this context, the main research contribution of this paper is three-fold. Firstly, it proposes a multi-modal data acquisition considering two different points of view for the cameras (i.e. lateral and frontal) and the inertial sensors placed on the wrist and on the fingers with the purpose to evaluate how the information on the fine movements can increase the recognition accuracy. Secondly, it proposes to classify a set of "basic" gestures and a set of daily living "scenes", as the composition of these "basic" gestures. Lastly, the dataset was acquired from the interaction with the Pepper robot to promote real-life similarities. Indeed, during the experimental session, the participants interacted with the robot and performed the actions that it was explaining and requiring. Particularly, 20 healthy subjects were enrolled and were requested to perform five daily living scenes (i.e. having lunch, personal hygiene, working, house cleaning, relax) in which they arbitrarily performed two or three "basic" gestures selected among 10 gestures which were similar in pair. In our previous work [14], we combined data from a depth camera mounted on a mobile platform, able to self-localize in

TABLE I

TABLE OF THE RELATED WORKS IN THE HUMAN ACTIVITY RECOGNITION FIELD: A SUMMARY OF THE SENSORS USED, THE PERFORMED ACTIVITIES, THE MACHINE LEARNING TECHNIQUES AND THE BEST PERFORMANCES.

Ref.	Used sensors[a]	Activities[b]	Machine Learning techniques	Accuracy
[6]	• **Wide** : cameras and inertial sensors on the wrist • **Fine**: -	• **Basic Gesture** : UTD-MHAD dataset [swipe left, swipe right, wave, clap, throw, arm cross, basketball shoot, draw X, draw circle (clockwise), draw circle (counter clockwise), draw triangle, walk, sit to stand, stand to sit, lunge, squat]. • **ADL**: UTD-MHAD dataset [bowling, boxing, baseball swing, tennis swing, arm curl, tennis serve, push, knock, catch, pickup and throw, jog]. • **Scene**: -	Collaborative representation classifier (CRC)	>0.97
[7]	• **Wide**: cameras and inertial sensors on the waist • **Fine**: -	• **Basic Gesture**: stand-to-sit, sit-to-stand, stand-to-lie, lie-to-stand, sit-to-lie, lie-to-sit, fall, waving a hand, flip to left, flip to right, counterclockwise rotation, clockwise rotation. • **ADL**: - • **Scene**: -	CNN and LSTM networks	Best result: 0.99
[8]	• **Wide**: smartphone inertial sensors in the pocket • **Fine**: -	• **Basic Gesture**: walking, upstairs, downstairs, sit, stand, lying. • **ADL**: - • **Scene**: -	Naive Bayes, KNN	Best result: 0.90
[9]	• **Wide**: cameras and inertial sensors on the wrist or thigh • **Fine**: -	• See activities Ref. [6]	K-nearest neighbors (KNN), SVM	Best result: 0.98
[10]	• **Wide**: smartphone inertial sensors in the pocket • **Fine**: -	• **Basic Gesture**: eight locomotion activities (still, walk, run, bike, car, bus, train, subway). • **ADL**: - • **Scene**: -	DT, RF, Naïve Bayes, KNN, SVM, Bagging, AdaBoost, XGB and MLP	Best result: 0.97
[11]	• **Wide**: inertial sensors on the wrist, chest and waist • **Fine**: -	• **Basic Gesture**: several types of falls. • **ADL**: cyclic activities of daily living (ADLs) e.g. walking and jogging and transient ADLs e.g. sitting down and lying down.	Multi-Layer Perceptron (MLP), SVM, KNN, RF, CNN, LSTM, SAE	Best result: 0.93
[12]	• **Wide**: cameras and smartphone inertial sensors in the pocket • **Fine**: - • **Scene**: -	• **Basic Gesture**: walking, sitting, standing, simulating tripping and falling down frontally. • **ADL**: walking and carrying an object, bending to pick up an object and coming back up, bending and staying down to tie shoelaces, drinking, picking up a phone call, bending to check under furniture and coming back up. • **Scene**: -	SVM, Ensemble classifier	SVM: 0.87, E: 0.91
[13]	• **Wide**: cameras and inertial sensors on the chest • **Fine**: -	• **Basic Gesture**: - • **ADL**: working on a laptop, watching TV, reading a book, operating a smartphone, vacuuming, lying in bed, preparing eggs, eating with the fork, washing the dishes. • **Scene**: -	KNN, Random Forest (RF)	>0.90 in the fusion approach
[14]	• **Wide**: cameras and inertial sensors on the wrist • **Fine**: inertial sensors on the fingers	• **Basic Gesture**: - • **ADL**: chop, drink with a glass, eat with a hand, eat with a spoon, open a pill container, talk on the phone, read a book, relax on the couch, stir, talk on the couch. • **Scene**: -	RF, SVM	0.77 in the best configuration
[15]	• **Wide**: cameras and inertial sensors on the wrist, neck, pocket of pants, waist and ankle • **Fine**: -	• **Basic Gesture**: walking, standing, picking up an object, sitting, jumping, laying and five human falls. • **ADL**: - • **Scene**: -	RF, SVM, MLP, kNN	95.93 ± 0.30 in the best configuration
[16]	• **Wide**: smartphone inertial sensors in the pocket • **Fine**: -	• **Basic Gesture**: inactive, active, walking, driving. • **ADL**: - • **Scene**: overall	SVM	67.22% ± 13.13%
This work	• **Wide**: cameras and inertial sensors on the wrist • **Fine**: inertial sensors on the fingers	• **Basic Gesture**: walking. • **ADL**: eat, drink, brush teeth, use laptop, write, talk on the phone, sweep, relax, read a book. • **Scene**: these activities were combined in 5 scene of daily living.	SVM, RF, KNN, LSTM	0.81 with feature-level fusion

[a] 'Wide' and 'Fine' indicate the sensors used to recognize the wide and fine movements, respectively. Particularly, with the term 'Wide' we refer to the whole body movements.

[b] As for the gesture mapping, the activities were clustered into 'Basic Gesture', e.g. walking, sitting, lying, 'ADL', related to a specific activity of daily living, and 'Scene', which includes all the activities composed by two or more activities without restrictions in the passage from one to the other.

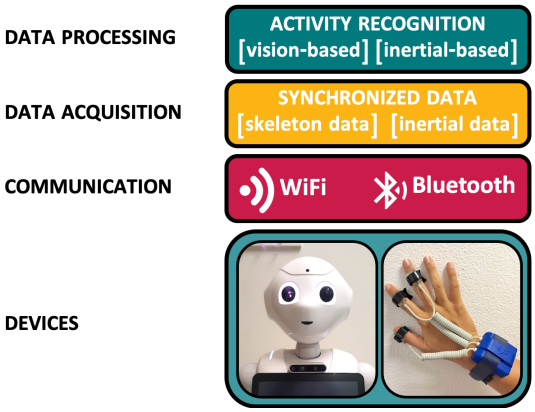


Fig. 1. System Architecture. On the bottom left: Pepper robot with the camera mounted on its chest; on the bottom right: SensHand.

Fig. 2. Experimental setup while performing the activity “Drink from a glass”.

the environment, and from a custom inertial wearable device named SensHand. However, in that case, an experimenter administered the test and we focused on the recognition of each single gesture. On the contrary, in this work, we focused also on the recognition of gestures as part of daily scenes and the Pepper robot autonomously addressed the protocol. Then, we evaluated the robustness of the proposed multi-modal system by using three machine learning algorithms (Support Vector Machine, Random forest and K-Nearest Neighbor), and one deep learning algorithm (Long Short-Time Memory). The use of well-known benchmark algorithms makes the results directly comparable with the previous works, thus the effectiveness of the proposed approach is more recognisable.

The operative objectives of this paper are two: i) to find the optimal combination of sensors that guarantees high operative accuracy in the recognition of gestures "hidden" into more complex scenes; ii) to investigate whether the system can manage the transitions from one activity to another by accurately classifying them.

III. Study Design

A. System description

As shown in Fig. 1, the proposed system is composed by:

- **Pepper Robot and Vision System**: Pepper robot is characterized by a multi-modal sensing (i.e. touch sensors, infrared, cameras and sonars) thanks to which it can interact with people and move in an autonomous way. In this paper, to enrich the visual capability of the robot, a RGB-D camera (i.e. Intel Realsense) was mounted on its chest over its tablet [20].
- **SensHand**: SensHand is composed of four customized inertial measurement units (IMUs) modules positioned on the wrist and on the intermediate phalanx of the thumb, index, and middle finger that are linked by spiral cables. Each module is composed of a complete 9-axis inertial sensor (6-axis geomagnetic module LSM303DLHC and 3-axis digital gyroscope L3G4200D, STMicroelectronics, Italy) and includes a microcontroller (ARM®-based 32-bit STM32F10RE MCU, STMicroelectronics, Italy) which can acquire, filter and store data at a frequency of 100 Hz. The wrist module is the coordinator of the device; it includes also a LiPo battery for power supply and a Bluetooth module for wireless data transmission. Moreover, it manages and synchronizes data coming from the fingers through the CAN-bus standard [21]. It is very easy to wear and to use thanks to its miniaturised and light structure; it is independent from the physical shape of the person wearing it.
- **Data Acquisition and Storage** The connection of the devices to a personal computer was established via Bluetooth for the SensHand and via WiFi for the robot. A Python interface was developed to manage the simultaneous acquisition of data and guarantee their correct storage. In particular, as concerns the inertial sensors, two Python executables were created: one to connect the device to the computer and the other to start the data acquisition and transmission. On the contrary, visual data have been acquired using the Robot Operating System (ROS) framework.

B. Participants

Twenty healthy participants of different ethnicity were enrolled for the experimentation, half males and half females, right-handed, from 19 to 44 years old. The experimental phase of this work was conducted in the Smart Interactive Technology (SIT) research laboratory of the Sheffield Hallam University (Sheffield, England, UK). At the beginning of the experimental session, written informed consent was obtained from the participants. As a token of gratitude, participants received an Amazon e-voucher of £10 after successfully completing the experiment. Study, design, and protocol, including subject privacy and sensitive data treatment, were approved by the Ethics Committee of the Sheffield Hallam University.

C. Experimental Protocol

The experimentation was designed to reproduce a real case scenario, in which the participants were asked to perform five different daily living scenes by alternating ten activities in the way they preferred. The selected five scenes were: Having Lunch (HL), Personal Hygiene (PH), House Cleaning (HC), Working (WO) and Relax (RE). Each scene was composed

of two or three "basic" gestures according to the mapping presented in Tab. II.

Before starting the test, the participant was asked to wear SensHand on the dominant hand (see Fig. 2). During the experimentation, Pepper robot gave instructions to the subject about how to perform the scene. If the participant did not understand, he could ask Pepper to repeat the assignment. Each scene was performed for one minute and the subjects could switch from one activity to another one as they preferred. They were also free to grab the objects and act in the way they preferred, so no instruction was given in that sense, giving the participants the possibility to act as naturally as possible. Pepper robot autonomously gave the start and the stop signals.

During the acquisition, each activity was labelled manually by an operator using the interface. The session was recorded by two cameras, one mounted over the robot and one located on the right side of the participant to acquire data from two different points of view. The lateral camera is the same as the one mounted on Pepper and it was placed at the same height from the ground. This double synchronous video acquisition has been used only in the experimental setting to save time, instead of asking the users to perform twice the protocol. It would not be necessary, therefore, in the standard use of the system. At the end of the experimentation, the users were asked to fill in the System Usability Scale (SUS) questionnaire to evaluate the usability of the wearable glove SensHand. A SUS score of 68 is considered usable, higher scores are considered above average [22].

IV. Gesture recognition

A. Feature Extraction

1) Camera: As concerns the RGB images analysis, skeleton features were obtained thanks to the Openpose software [23]. In particular, from each frame, 25 keypoints were estimated for the body, where each of them represents the (x, y) pixels' coordinates of the joints. In this study, we considered only a restricted set of joints, composed by: head, neck, hands, feet and torso, which are among the most discriminant information for activity recognition [24].

The extracted features have then been normalized moving the original reference frame from the camera to the torso joint, and scaling the joints with respect to the distance between the neck and the torso [24], [14]. This results in a set of data which are independent with respect to the person's specific size and to the relative position of the camera. The posture feature vector was composed by 12 attributes, which corresponded to the x and y coordinates of the restricted set of joints, excluding the torso which was used as reference. The signal containing the skeleton features for each frame was segmented by 50 %-overlapping moving windows with a size of 3 s, and for each window the mean x and y joints' coordinates were extracted. This procedure was performed for the frontal (FC) and the lateral (LC) cameras.

2) Inertial Data: According to the results obtained in [25], in this study only the data coming from the wrist and index finger sensors of SensHand were used. A Fourier analysis of the raw signal was performed to identify the cut-off frequency. Since the main frequencies of the signal were between 0 and 4, a 4th order digital low-pass Butterworth filter was used setting the cut-off frequency at 5 Hz. In particular, acceleration and angular velocity data were first filtered on their single components (x, y, z), and then concatenated computing the Euclidean norm. These data were then segmented in 3 s windows, same as skeleton data, and, from each of them, different features were extracted. The final dataset was composed by 10 features related to acceleration values, i.e. mean, standard deviation, variance, mean absolute deviation (MAD), root mean square (RMS), skewness, kurtosis, signal magnitude area (SMA), normalized jerk and power, and 6 features to angular velocities, i.e. mean value, standard deviation, variance, MAD, RMS and power. These features were computed for both wrist (W) and index finger (I), for a total of 32 features. In the experimentation, we made a continuous acquisition of data without interruption between activities for each scene. Therefore, the resulting database has two data types: (i) "pure data", which are clearly related to only one of the activities in a scene; (ii) "transition data", that catch the transitions from one activity to another. In the analysis presented in this article, when segmenting the data in the window of three seconds, if two (or more) were annotated as "transition", the sequence was labeled as corresponding to the most frequent activity.

B. Classification

At the end of the features extraction, the Kruskal Wallis test was applied to obtain the most significant feature vector in distinguishing the group of instances. This test confirmed that the ten gestures, which characterized the activities under investigation, were statistically different for all the extracted features ($p < 0.05$). Then, a correlation analysis was performed in order to retain only the significantly uncorrelated features (Correlation Coefficient < 0.85), as in [25], and to use them as input for the machine learning algorithms.

The system was evaluated by considering both inertial and visual sensors as stand-alone systems and by fusing the two sensor modalities at feature-level [26] for a total of eleven datasets: frontal camera (FC), lateral camera (LC), index finger (I), wrist (W) and their combination (IW, I+FC, I+LC, W+FC, W+LC, IW+FC, IW+LC).

In the literature, several methods are used for the classification and recognition of human gestures [27] (Table I). Related works employed supervised machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP) and k-Nearest Neighbors (kNN). More recently, traditional machine learning techniques are compared to the deep learning algorithms. Among them, convolutional neural network (CNN) and recurrent neural network (RNN) are two popular ones [11]. Therefore, in this paper, three commonly used supervised machine learning algorithms and one deep learning approach were employed in the stand-alone and in the combined classifications:

- Multiclass Support Vector Machine (SVM): it takes data which are not linearly separable in the input space and turns them into linearly separable data in the higher dimensional feature space. Then, it finds the hyperplane that can separate the classes with the largest margin. A third order polynomial kernel function has been used in

TABLE II
DESCRIPTION OF GESTURES PERFORMED DURING THE EXPERIMENTAL SESSION. IN THE LAST COLUMN, SR STANDS FOR "SITTING ON THE CHAIR", ST FOR "STANDING" AND SC FOR "SITTING ON THE COUCH".

Activity	Description	Scene	Position
EF: Eat with the fork	Take the fork from the table, eat and put the fork back	Having Lunch (HL)	SR
DG: Drink from a glass	Take a glass from the table, drink and put it back	Having Lunch (HL), Personal Hygiene (PH)	SR
BT: Brush teeth	Take the toothbrush, brush teeth and put it back	Personal Hygiene (PH)	SR
UL: Use laptop	Type on the keyboard with both hands	Working (WO)	SR
WP: Write on a paper	Take a pen and write on a paper	Working (WO)	SR
TP: Talk on the phone	Take the phone, talk on it and put it back	Working (WO), Relax (RE)	SR
WK: Walk	Walk forward and backward	House Cleaning (HC)	ST
SB: Sweep with the broom	Take the broom, sweep and put it back at the end	House Cleaning (HC)	ST
RC: Relax on the couch	Sit comfortably on the couch and relax	Relax (RE)	SC
RB: Read a book	Take the book, read it and turn pages	Relax (RE)	SC

this work.

- Random Forest (RF): it operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees, with the goal of reducing the variance.
- K-Nearest Neighbor (KNN): it stores all available data and classifies new cases based on similarity measures, which are distance functions. They are assigned to the most common class among its k nearest neighbors. We set k = 1, so the object was simply assigned to the class of that single nearest neighbor, and the euclidean distance was used as distance metric.
- Long Short-Term Memory (LSTM): it is a recurrent neural network (RNN) employed in the field of deep learning. The proposed LSTM has batch size equal to 8, 512 hidden layers and 128 fully connected layers. We use the Adam optimizer (learning rate = 0.001) and L2 regularization (preventing overfitting) during training. Also, we applied cross-Entropy as loss Function. Three different number of epochs were used to train the model (i.e. 100, 200 and 300). For this classifiers we use all the selected features that were further normalized in the range (-1,+1) before applying the LSTM classification.

These datasets were classified using a 10-fold cross-validation technique. The final classification results are obtained as average of the performances of the ten created models.

C. Evaluation

All the analyses were performed using Matlab2020a and Pytorch for machine learning and deep learning, respectively. The classification performances were evaluated in terms of accuracy, precision, recall and F-measure, which can be described as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. The classification time was also computed by a built-in MATLAB function, a stopwatch counter named *tic*, which measures the amount of time that the classifiers take to complete each classification step. In particular, it has been computed for every classifier while performing the cross-validation and for every combination of sensors.

V. RESULTS

In this section, the results obtained from the multiple comparisons of the four supervised classifiers on eleven datasets are reported in detail. Overall, 3361 windows were created for each combination of data concerning the frontal camera (FC, I+FC, W+FC, IW+FC), while 3213 for the ones concerning the lateral camera (LC, I+LC, W+LC, IW+LC). Moreover, 3213 windows were considered for inertial sensors alone (I, W, IW). The number of rows corresponds to the number of windows in each dataset. On the contrary, the number of columns varies for each combination and depends on the number of features. After the feature selection process, the inertial features shown in Tab. III were retained for the machine learning classification step. As regards skeleton features, all of them were selected and used by the classifiers. For what concerns LSTM approach, all the features were used as input. In this paper, we only report the LSTM classifier's best performances (300 epochs), while the other results (100 and 200 epochs) are reported in the supplementary material.

Generally, the results obtained under the multi-modal datasets are better than those achieved when considering the sensors separately and they are comparable with the related works (Tab. I). As concern the stand-alone configurations, the results in Tab. IV show that the system composed by the inertial sensors on the index finger (I) and on the wrist

TABLE III
INERTIAL FEATURES SELECTED AFTER CORRELATION ANALYSIS.

Index+Wrist		Index/Wrist
Wrist acc. mean	Index acc. mean	Acc. mean
Wrist acc. stdev	Index acc. stdev	Acc. stdev
Wrist acc. RMS	Index acc. RMS	Acc. RMS
Wrist acc. skewness	Index acc. skewness	Acc. skewness
Wrist acc. kurtosis	Index acc. kurtosis	Acc. kurtosis
Wrist acc. SMA	Index acc. SMA	Acc. SMA
Wrist acc. power	Index acc. power	Acc. power
Wrist ang. vel. mean	Index vel. mean	Ang.vel. mean
Wrist ang. vel. stdev	Index vel. power	Ang.vel. stdev
Wrist ang. vel. power		Ang.vel. power

TABLE IV
RESULTS OBTAINED BY STAND-ALONE SYSTEMS.

	Accuracy	Recall	F-meas	Precision	Time [s]
Index (I)					
SVM	0.52	0.50	0.51	0.52	88.61
RF	0.55	0.53	0.54	0.56	16.62
KNN	0.51	0.49	0.50	0.52	26.69
LSTM	0.56	0.56	0.56	0.57	4897.21
Wrist (W)					
SVM	0.54	0.52	0.53	0.53	88.95
RF	0.57	0.55	0.56	0.58	14.88
KNN	0.50	0.48	0.49	0.50	28.02
LSTM	0.57	0.55	0.55	0.56	3283.57
I+W					
SVM	0.65	0.64	0.64	0.65	81.25
RF	0.64	0.63	0.63	0.64	21.90
KNN	0.61	0.59	0.60	0.62	27.72
LSTM	0.66	0.65	0.65	0.66	4285.47
FC					
SVM	0.77	0.77	0.77	0.77	97.50
RF	0.75	0.75	0.75	0.76	16.92
KNN	0.75	0.76	0.76	0.77	30.14
LSTM	0.68	0.69	0.69	0.71	5493.14
LC					
SVM	0.69	0.70	0.70	0.72	21.47
RF	0.68	0.68	0.69	0.71	17.08
KNN	0.69	0.70	0.71	0.73	30.23
LSTM	0.61	0.62	0.62	0.65	3609.12

(W) obtains accuracy levels up to 0.56 and 0.57, respectively, while 0.65 of accuracy is obtained when considering the I+W combination. The frontal camera (FC), which has a good view of the user performing the activity, is able to recognize the gestures with 0.77 of accuracy, recall, F-measure and precision, with the SVM classifier. These values decrease when considering the camera positioned on the side (LC). In this case, the accuracy, recall, F-measure and precision are 0.69, 0.70, 0.71 and 0.73, respectively, with KNN classifier. RNN and machine learning had comparable performances, even if the LSTM required longer training time.

As concern the multi-modal datasets, the obtained results show that the fusion-at-feature-level approach improves the classification accuracy compared to the use of the independent classifiers. Similarly to the stand-alone combinations, deep learning and machine learning showed comparable performances. Also in this case, LSTM had a longer training time with respect to the other algorithms. Tab. V indicates that the system is able to recognize the ten activities with 0.81 (IW+FC) and 0.75 (W+LC) of accuracy as best configurations, obtained with SVM and RF, respectively. Fig. 3 compares the F-measure values obtained by these two combinations with the other ones for every single activity. I+FC, W+FC and IW+FC obtained comparable performance in classifying the 10 gestures; similar results were obtained also with the I+LC, W+LC and IW+LC datasets. From a visual inspection it is clear that the worst configuration is IW (blue line) and that the use of multi-sensors approach increases the performance of the system. As for the FC dataset classified with SVM (the best stand-alone configuration), the system can correctly classify almost all the activities with an average F-Measure equal to 0.77, with some difficulties in recognizing the activities “Drink from a glass (DG)” (0.7), “Sweep with the broom (SB)” (0.7) and “Walk (WK)” (0.6). As expected, it is evident from a visual inspection that the results from the LC are worse than the ones obtained with the FC dataset, but also from the ones obtained with the IW dataset for BT and WK gestures. Indeed, the use of combined sensors increases the recognition performances also when the robot is not in the optimal position with respect to the user. Indeed, with respect to the only LC, the I+LC increases the overall accuracy by 0.6 with the best configuration (obtained with RF). In particular, by looking at the single activities, the F-measure values increase by 0.2 for the activity “Brush teeth (BT)” and by 0.1 for the activity “Write on a paper (WP)”.

As concerns the evaluation of the usability of SensHand, the results show that the average SUS score is 72, the standard deviation 14.5, the maximum 95, and the minimum 40.79. The results underline a Good Usability of SensHand (Letter Grade B) [28].

VI. DISCUSSION

In this work, cameras and wearable inertial sensors have been combined to enhance the capabilities of the robot to recognize human activities. Indeed, the results make clear that the inertial sensors alone are not enough to recognize all the ten activities in a consistent way, since the levels of accuracy, recall, F-measure and precision are insufficient for a reliable application, i.e. around 0.65 (IW). Better results are achieved considering the FC (0.77 of average accuracy). However, this accuracy value significantly decreases when the camera is in the lateral position (0.69).

Comparing these results with those relating to previous works, some of them reach a slightly higher level of accuracy; however, they classify "basic" activities that are very different from each other (e.g. walking, sitting, pick-up and object), and they did not include data relating to the finger movements, nor those relating to the "scene" (see Table I). For this reason, we cannot make a direct comparison between these works. Furthermore, we must take into account two aspects

TABLE V
FUSION AT FEATURE-LEVEL'S RESULTS. "A" STANDS FOR ACCURACY, "R" FOR RECALL, "F" FOR F-MEASURE, "P" FOR PRECISION AND "T" FOR CLASSIFICATION TIME EXPRESSED IN SECOND.

	I+FC					W+FC					IW+FC				
	A	R	F	P	T	A	R	F	P	T	A	R	F	P	T
SVM	0.79	0.79	0.79	0.79	207.0	0.79	0.79	0.79	0.79	188.26	0.81	0.80	0.81	0.81	356.20
RF	0.77	0.77	0.77	0.78	24.69	0.77	0.76	0.77	0.78	24.48	0.77	0.77	0.77	0.78	28.36
KNN	0.74	0.73	0.74	0.75	36.58	0.73	0.72	0.73	0.75	30.92	0.76	0.75	0.76	0.76	33.84
LSTM	0.74	0.75	0.75	0.76	5475.76	0.74	0.75	0.75	0.75	3437,64	0.74	0.75	0.74	0.76	4473.61
	I+LC					**W+LC**					**IW+LC**				
	A	R	F	P	T	A	R	F	P	T	A	R	F	P	T
SVM	0.72	0.72	0.73	0.75	140.63	0.72	0.71	0.72	0.74	133.30	0.74	0.74	0.75	0.77	274.1
RF	0.75	0.74	0.75	0.77	25.04	0.75	0.75	0.76	0.77	21.20	0.74	0.74	0.74	0.75	25.58
KNN	0.65	0.65	0.66	0.68	42.77	0.64	0.64	0.65	0.67	29.53	0.71	0.65	0.67	0.70	32.3
LSTM	0.72	0.73	0.73	0.73	5239.81	0.70	0.71	0.71	0.72	3287.42	0.73	0.74	0.74	0.74	4310.43

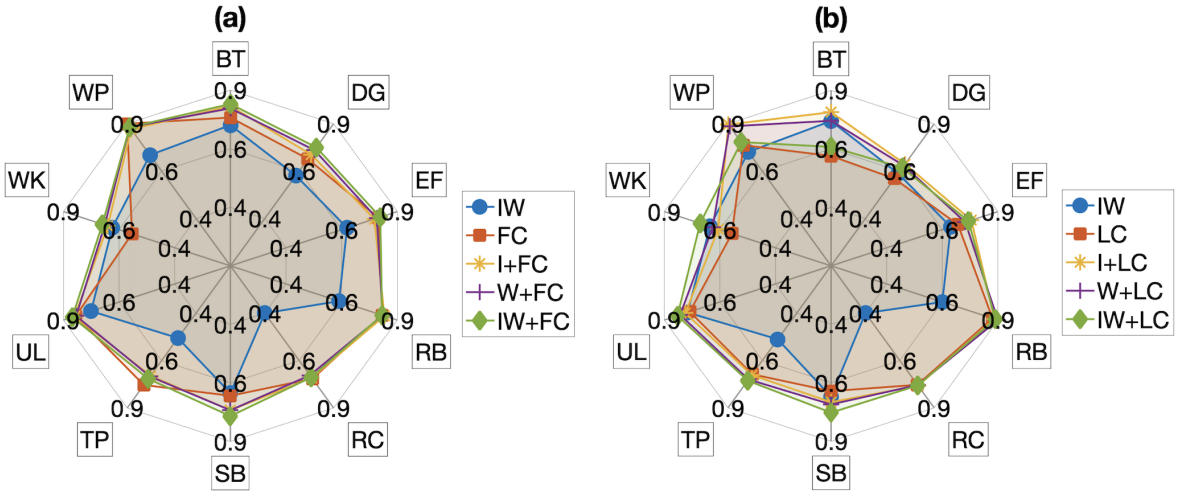


Fig. 3. Spider plots in which F-measure values on the axes are compared across the ten activities with different combination of sensors. In particular, frontal and lateral camera are considered in (a) and (b), respectively. This figure reports only the best machine learning classifiers for each configuration: SVM for FC, IW, I+FC, W+FC, IW+FC, IW+LC, RF for I+LC, W+LC and KNN for LC.

related to these results: first, in a life-like situation, the robot is unlikely to be positioned exactly in front of the person performing the activity, but it will be more likely in a non optimal position for the recognition, e.g. on a side. In order to evaluate more realistic situations, this paper studied two different visual perspectives (frontal and lateral) to explore how the relative position between the robot and the user could affect the recognition task. We found out that the visual recognition system decreases its capabilities by moving the position of the camera by only 90 degrees, i.e. it lost 8% of accuracy when the camera was in a lateral position. This is due to the natural occlusion problems that make more difficult to recognize gestures from the video of a single camera. However, other issues can reduce the performance also when the camera is frontally. Fig. 3 shows that, even if the overall average accuracy of the system is quite high, some activities like "Sweep with the broom (SB)" and "Walk (WK)" are not well recognized by the system, with a recognition performance significantly below the average. This is because the way arms and hands move is different in the two activities, while the legs act in the same way, so it may be difficult for a camera to recognize this slight difference.

These issues can be overcome by fusing the information that is acquired from the cameras, able to capture the gross motor actions of the body, with the inertial wearable device, able to capture the fine movements of the hand. Indeed, four different configurations have been tested with a feature-level fusion approach, i.e. features from the frontal camera, wrist and index (IW+FC), from frontal camera and wrist (I+FC), from the lateral camera, wrist and index finger (IW+LC) and from the lateral camera and the index finger (I+LC), to understand which is the best combination of sensors.

As shown in Tab. V, the fusion of inertial features with the frontal camera leads to the best results achieved in this study. Indeed, not only the overall accuracy of the system is the highest on average, but the recognition performance of almost

every single activity is better than the standalone configuration (Fig. 3).

Furthermore, it is important to remark that the paper aims to reproduce real operative conditions as far as possible. Indeed, in each scene, the participants could switch freely from one activity to another in one minute. In this amount of time, the cameras and the SensHand recorded visual and inertial data continuously, therefore the acquisition stream was unique for each scene. For this reason, the novelty with respect to Manzi et al. [24], whose system achieved 0.77 of accuracy, is that in our work the data corresponding to the transitions from one activity to the other were present in the acquisition signal, and the system revealed to be good enough in classifying them without losing much in terms of performances (0.81 of accuracy). Results in Fig. 3 suggest that it is still possible to obtain very good performances by using the robot's camera combined with only one inertial sensor on the user's hand, which performs even better in activities like "Brush Teeth" (BT), "Drink from a glass" (DG), "Eat with the fork"(EF) and "Write on a paper" (WP) in the lateral configuration. On the contrary, the activities "Relax on the couch" (RC) and "Talk on the phone" (TP) are better recognized by the frontal camera alone. This aspect may be discussed by considering that these two activities are quite static, therefore the inertial sensors do not give any additional information about the hand, but, conversely, it introduces some errors. It is worth remarking that, differently from other public datasets with activities of daily living, our dataset was created by choosing gestures that were similar in pairs, adding extra difficulty to the system and taking a step forward to recognize them. It is important also to remark that by looking at every single activity, the index sensor combined with the camera performs better than the wrist one, so the final system should be lighter and easier to wear. In this sense, a smart ring could be a good trade-off between comfort and high performance in recognizing gestures. Additionally, in this paper, we evaluated the performances of the system by comparing different classification algorithms, i.e. from traditional machine learning to innovative deep learning (LSTM). The results show that the latter does not produce the best performances. As reported in literature [29], this is reasonable, since the machine learning methods perform better than deep learning with small data size, as it is in our case.

For what concerns the limitations of this study, it is still important to remark that this analysis has been conducted by classifying windows of three seconds of the signal. However, in real operative scenarios, the robot will not capture and analyse the data coming from a single event-window but it will observe the person for a longer period of time. Future studies could take advantage of this issue by considering the event-window as a part of the event's flow, so that it could be easier to analyse the recognition performances when the algorithms also consider the previous and the consecutive event. Future studies can create ensemble techniques that combine multiple machine learning algorithms, including Deep Learning, by merging the predictions of different classifiers at the decision level to increase the overall accuracy. Particularly, a hybrid approach could be investigated to understand how to combine vision and inertial raw data with Deep Learning approach. Finally, the participant groups could be extended including people of different ages, thus to investigate whether the recognition can differ with age.

VII. Conclusion

Gesture recognition is a crucial aspect to consider to improve human-robot cooperation. In this article, we proposed a multi-modal system, composed of a Pepper robot and an inertial device, named SensHand, which was tested in a realistic environment by 20 healthy participants who were instructed by the robot to execute 10 different gestures. In summary, the results indicate that a social robot with an embedded camera could recognize people's gestures in a realistic scenario, and that the system recognition capabilities can be stronger if the person wears the inertial sensors on his dominant hand while performing daily living activities. Indeed, the cameras could capture the big movements of the body, whereas the inertial sensors can catch fine movements of the hand, especially when they are not in the camera's field of view. Additionally, these findings suggest that the multi-modal sensor approach could improve the recognition even when the robot is not positioned in front of the human, improving the task recognition and the human-robot cooperation.

Acknowledgment

The authors would like to thank all the volunteers that participated to our experimental session.

References

[1] O. Nocentini, L. Fiorini, G. Acerbi, A. Sorrentino, G. Mancioppi, and F. Cavallo, "A survey of behavioral models for social robots," *Robotics*, vol. 8, no. 3, p. 54, 2019.

[2] S. C. Akkaladevi and C. Heindl, "Action recognition for human robot interaction in industrial applications," in *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, 2015, pp. 94–99.

[3] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[4] H. Liu and L. Wang, "Gesture recognition for human-robot collaboration: A review," *International Journal of Industrial Ergonomics*, vol. 68, pp. 355–367, 2018.

[5] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, 12 2015.

[6] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2016.

[7] N. Dawar and N. Kehtarnavaz, "Action Detection and Recognition in Continuous Action Streams by Deep Learning-Based Sensing Fusion," *IEEE Sensors Journal*, vol. 18, no. 23, pp. 9660–9668, 2018.

[8] A. Wang, G. Chen, J. Yang, S. Zhao, and C. Y. Chang, "A Comparative Study on Human Activity Recognition Using Inertial Sensors in a Smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.

[9] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust Human Activity Recognition Using Multimodal Feature-Level Fusion," *IEEE Access*, vol. 7, pp. 60 736–60 751, 2019.

[10] M. Gjoreski, V. Janko, G. Slapničar, M. Mlakar, N. Reščič, J. Bizjak, V. Drobnič, M. Marinko, N. Mlakar, M. Lustrek, and M. Gams, "Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors," *Information Fusion*, vol. 62, 04 2020.

[11] M. Saleh, M. Abbas, and R. B. Le Jeannès, "Fallalld: An open dataset of human falls and activities of daily living for classical and deep learning applications," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1849–1858, 2021.

[12] H. Li, A. Shrestha, F. Fioranelli, J. Le Kernec, H. Heidari, M. Pepa, E. Cippitelli, E. Gambi, and S. Spinsante, "Multisensor data fusion for human activities classification and fall detection," *Proceedings of IEEE Sensors*, vol. 2017-Decem, pp. 1–3, 2017.

[13] P. Voigt, M. Budde, E. Pescara, M. Fujimoto, K. Yasumoto, and M. Beigl, "Feasibility of human activity recognition using wearable depth cameras," *Proceedings - International Symposium on Wearable Computers, ISWC*, pp. 92–95, 2018.

[14] A. Manzi, A. Moschetti, R. Limosani, L. Fiorini, and F. Cavallo, "Enhancing Activity Recognition of Self-Localized Robot Through Depth Camera and Wearable Sensors," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9324–9331, 2018.

[15] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "Up-fall detection dataset: A multimodal approach," *Sensors (Switzerland)*, vol. 19, no. 9, 2019.

[16] D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, and M. R. Luaces, "A public domain dataset for real-life human activity recognition using smartphone sensors," *Sensors*, vol. 20, no. 8, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/8/2200

[17] H. Cheng, L. Yang, and Z. Liu, "Survey on 3d hand gesture recognition," *IEEE transactions on circuits and systems for video technology*, vol. 26, no. 9, pp. 1659–1673, 2015.

[18] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[19] J. Qi, P. Yang, A. Waraich, Z. Deng, Y. Zhao, and Y. Yang, "Examining sensor-based physical activity recognition and monitoring for healthcare using internet of things: A systematic review," *Journal of biomedical informatics*, vol. 87, pp. 138–153, 2018.

[20] "Pepper, soft bank robotics," https://www.softbankrobotics.com/emea/en/pepper.

[21] F. Cavallo, A. Moschetti, D. Esposito, C. Maremmani, and E. Rovini, "Upper limb motor pre-clinical assessment in parkinson's disease using machine learning," *Parkinsonism & related disorders*, vol. 63, pp. 111–116, 2019.

[22] J. Brooke, "Sus: a "quick and dirty'usability," *Usability evaluation in industry*, vol. 189, 1996.

[23] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[24] A. Manzi, P. Dario, and F. Cavallo, "A human activity recognition system based on dynamic clustering of skeleton data," *Sensors (Switzerland)*, vol. 17, no. 5, 2017.

[25] A. Moschetti, L. Fiorini, D. Esposito, P. Dario, and F. Cavallo, "Recognition of daily gestures with wearable inertial rings and bracelets," *Sensors (Switzerland)*, vol. 16, no. 8, 2016.

[26] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2015.00028

[27] E. De-La-Hoz-Franco, P. Ariza-Colpas, J. M. Quero, and M. Espinilla, "Sensor-based datasets for human activity recognition – a systematic review of literature," *IEEE Access*, vol. 6, pp. 59 192–59 210, 2018.

[28] "System usability scale (sus) grade," https://www.usabilitest.com/system-usability-scale.

[29] Y. Pu, D. B. Apel, V. Liu, and H. Mitri, "Machine learning methods for rockburst prediction-state-of-the-art review," *International Journal of Mining Science and Technology*, vol. 29, no. 4, pp. 565–570, 2019, sI: Recent Advancements in Mine Safety Science and Engineering. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095268619302812